INTRODUCTION AU STOCKAGE

MAJEURE INFRASTRUCTURE 2020

Formation Ynov

Florent Paillot

Présentation du cours

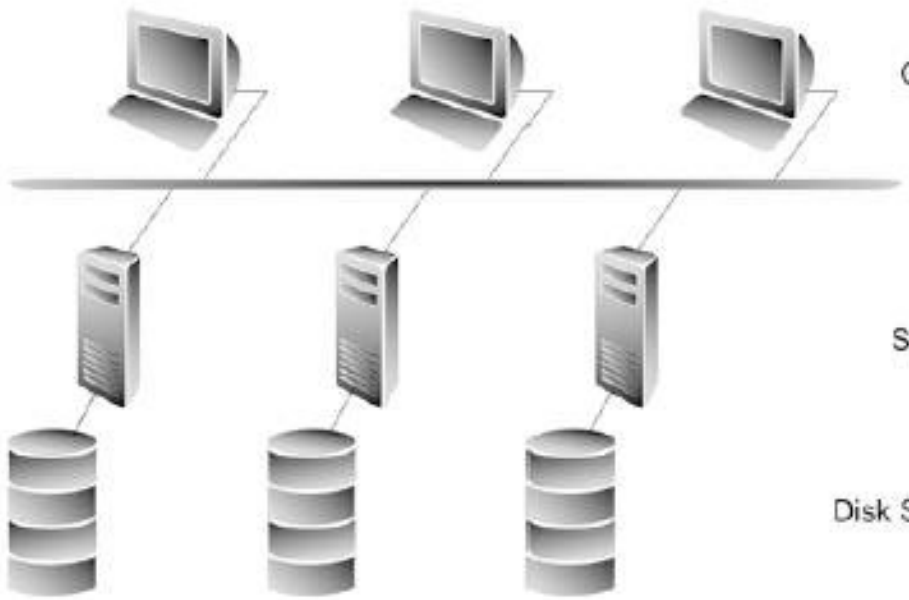
- I) Le stockage direct VS stockage en réseau
- II) Périphériques de stockage et sous systèmes
- III) Gestion de la redondance des données
- IV) Les systèmes de fichier

I - Le stockage réseau

- DAS et SAN
- Les fonctions du stockage réseau
- Le chemin des I/O

I – Les DAS

- Direct Attached Storage



DAS limitations:

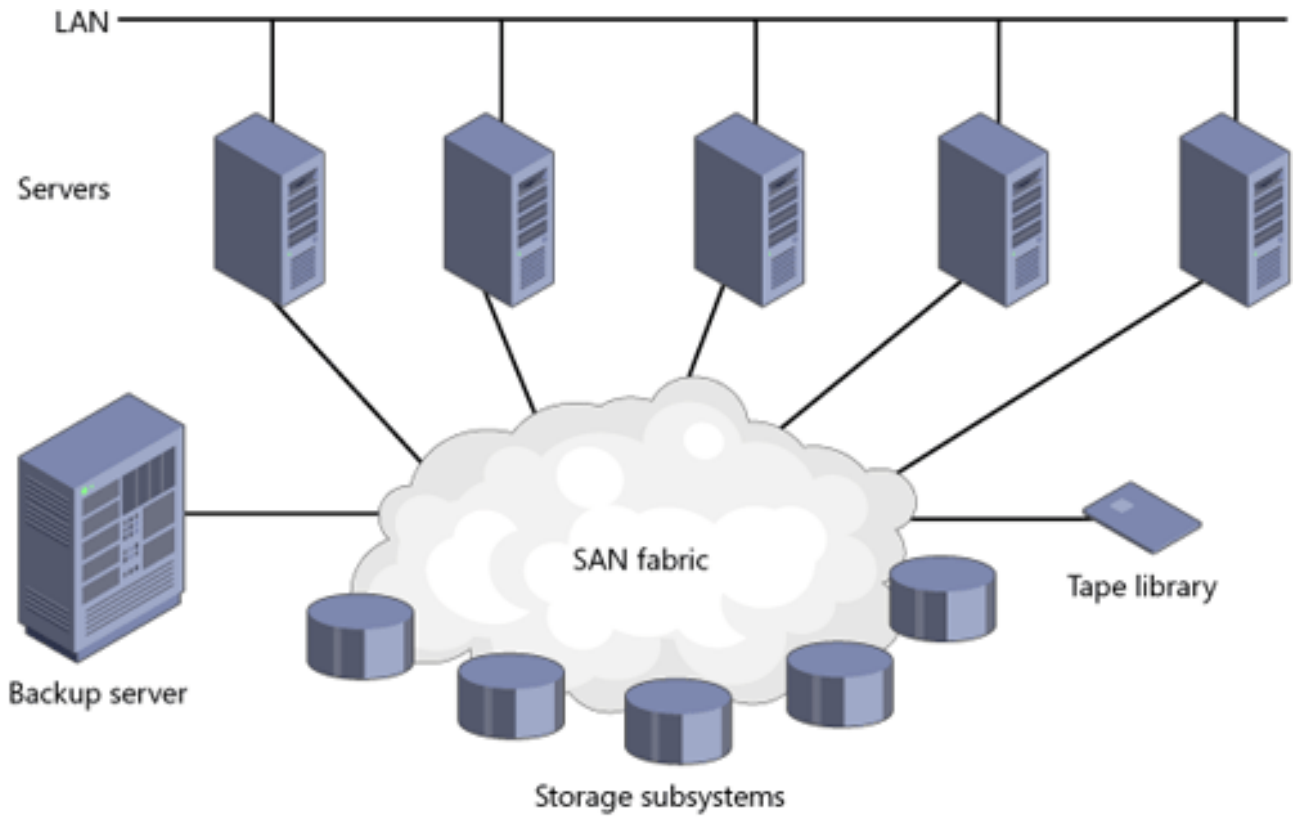
- Difficult to manage
- Poor scaling
- Poor capacity utilization
- Limited availability

I – Les DAS

- Système historique pour la mise à disposition de fichiers
- Relation de type « one server to many storage »
- Problèmes :
  - De disponibilité
  - D’évolutivité
  - De distance
  - De sous utilisation

I – Les SAN

- Storage Area Networks



I – Les SAN

- Relation de type « many to many »
- Permet de lever les limitations des DAS et d’envisager un accès différent aux données

I – Les fonctions du stockage réseau

Pour mieux comprendre les réseaux de stockage, on peut analyser leurs composants en les classant dans trois grandes catégories, proposées par Marc Farley :

Fonction	Description
Connecter	Transmission de données entre les systèmes et le stockage
Stocker	Applications bas niveau utilisant des commandes spécialisées et des protocoles de contrôle pour les interactions systèmes/périphériques
Remplir	Dirige le placement des objets de stockage et est responsable de la présentation des données aux applications et aux utilisateurs

I – Les fonctions du stockage réseau

- Connecter :
  - Les bus (principalement en DAS)
    - SCSI, ATA, ..
    - Fournissent une garantie dans l'ordre de livraison
    - Limités en longueur/performances
  - Les réseaux
    - FCP, iSCSI, FC/IP, ..
    - Fournissent des capacités de gestion plus aboutis
    - Sont plus complexes à mettre en place/maintenir

I – Les fonctions du stockage réseau

- Stocker

- Les protocoles de stockage sont orientés commande/réponse, on retrouve un couple initiateur (initiator) / cible (target)

- On parle de blocs d'I/O, ils sont composés :
  - D'un type d'opérations à réaliser
  - D'une adresse sur laquelle intervenir
  - De données de stockage ou de contrôle

- Les contrôleurs de stockage sont présents dans :
  - Les **H**ost **B**us **A**dapter
  - Les Périphériques de stockages
  - Les sous systèmes de stockage
  - Certains switchs dédiés au stockage

I – Les fonctions du stockage réseau

- Stocker

- Les disques durs, les disques virtuels, les LUN possèdent fondamentalement le même rôle. C’est leur assemblage qui permet de rendre le service.

- L’adressage dans le monde du stockage :
- Schéma

I – Les fonctions du stockage réseau

- Remplir

- L’objectif est d’organiser la façon dont sont structurées les données dans un espace d’adressage de stockage et de les présenter aux applications et aux utilisateurs.

- Deux types de produits remplissent ce rôle : les systèmes de fichiers et les bases de données

- Les composants chargés de remplir, sont aussi chargés de gérer une partie de la sécurité via des ACL

I – Les fonctions du stockage réseau

- Les applications savent comment sont structurées leurs propres données mais ont besoin d'abstraction vis-à-vis des couches plus basses.

- Le système de fichier ne sait pas comment les applications structurent leur données mais leur propose un support normalisé pour accéder à l'espace d'adressage de stockage : les I/O de type fichier.

- Une application peut demander au SF les 50 premiers octets d'un fichier sans avoir à connaître la structure sous-jacente

I – Les fonctions du stockage réseau

- La différence principale entre les NAS et les SAN se situe au niveau des fonctions qu'ils fournissent :
  - Les SAN interviennent sur la fonction « stocker »
  - Les NAS interviennent sur la fonction « remplir »

- Exemple de chaîne complète
  - Schéma

I – Chemin des I/O

- Les besoins pour le transport d'I/O de stockage :
  - Performance :
    - Bande passante
    - Latence
    - Fiabilité
    - Intégrité (ordre de livraison)

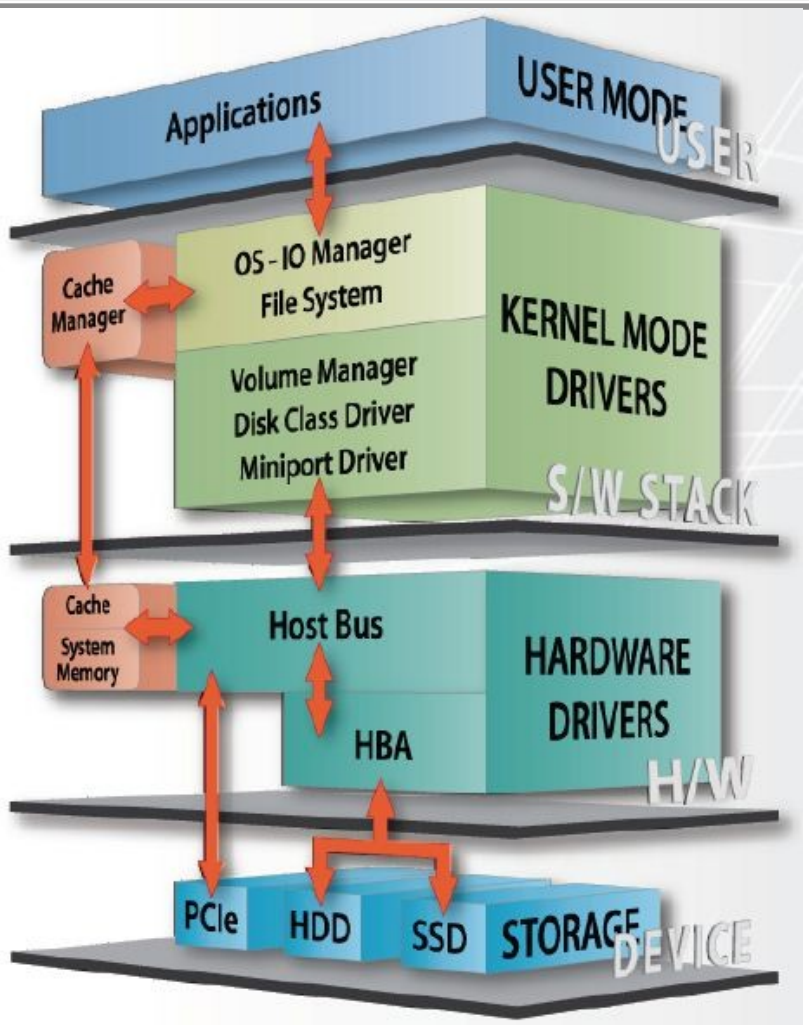
I – Chemin des I/O

- Lors du transport des I/O on retrouve :

Component	Hardware	Software
Host system components	Processor Memory Memory bus I/O bus Network interface	Applications File system Volume manager MP diver HBA storage driver
Network components	Network media Port hardware Buffer systems Switch core	Storage application Routing logic Flow control Virtual network
Storage susbsystem components	Network interface Storage controller Cache memory Subsystem bus/network Storage device Storage media	Target/logical unit number (LUN) mapping Virtual disk management

I – Chemin des I/O

- Interactions entre les différentes couches

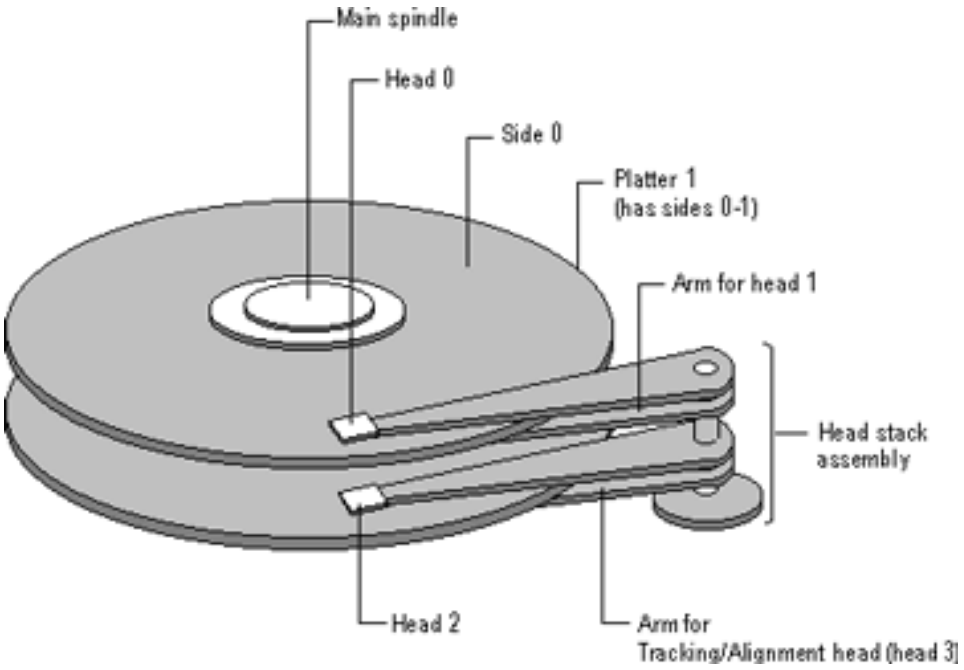


II - Périphériques de stockage et sous systèmes

- Les périphériques de stockage
- Les sous systèmes
- SCSI
- Les types de connexions

II – Les disques durs

- Les données sont stockées sur des plateaux entraînés par des axes à vitesse constante,



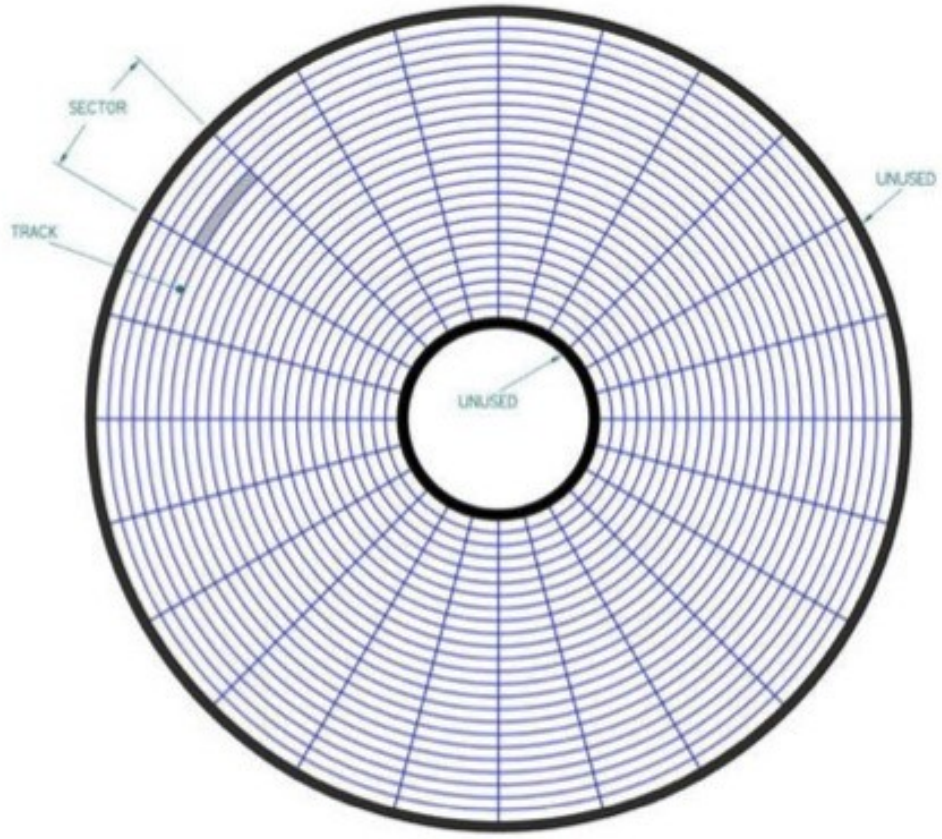
II – Les disques durs

- Apparition en 1956 (RAMAC 305) : 5Mo

- Le secteur est la plus petite unité d'allocation, en général 512 octets ou 4Ko (disques récents)

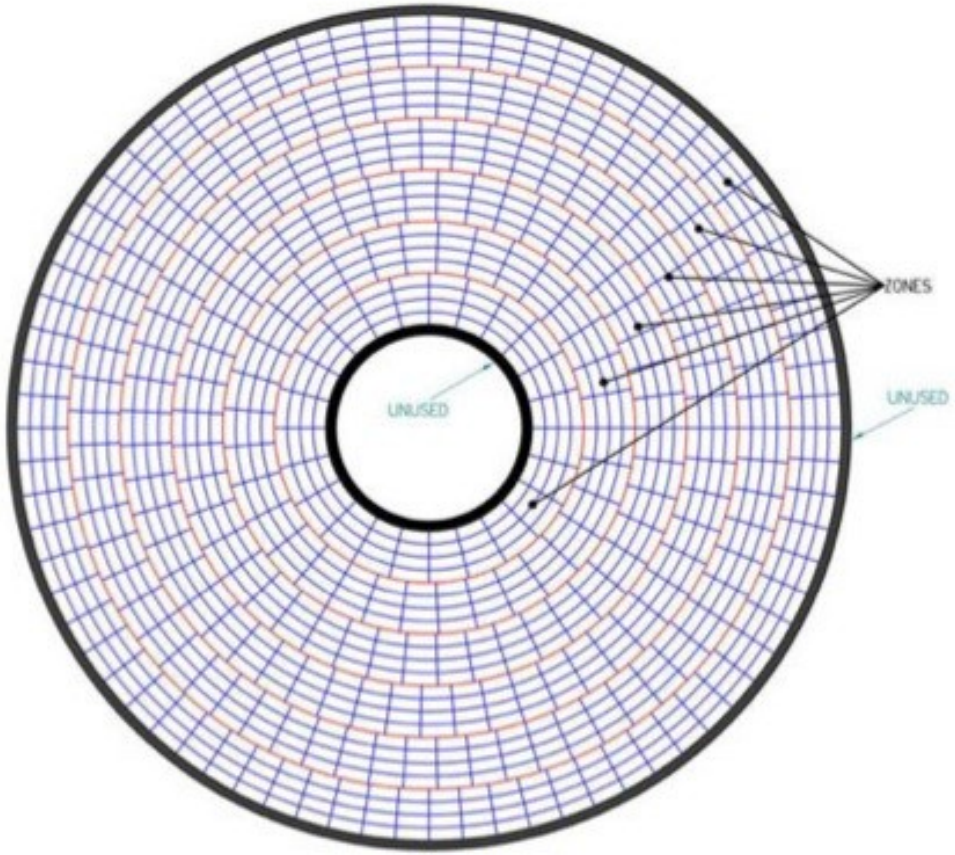
II – Les disques durs

Structure géométrique du disque



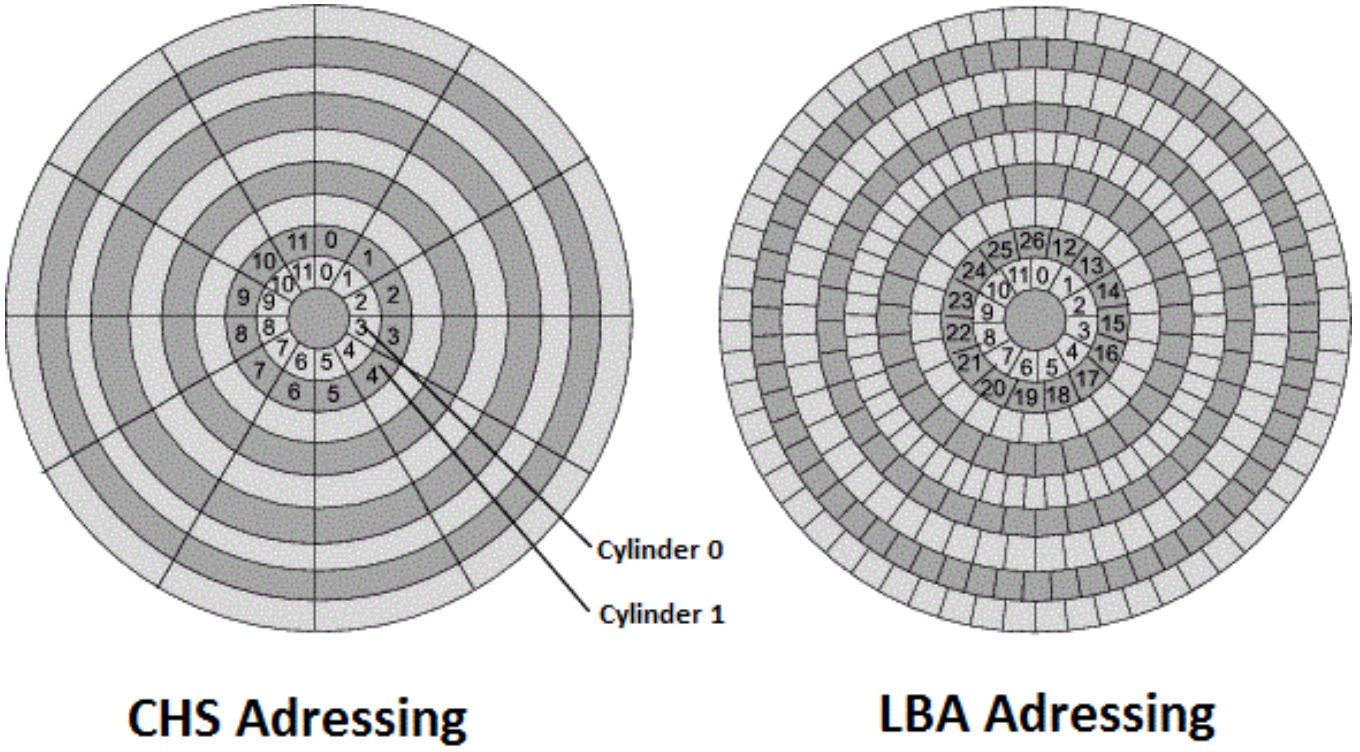
II – Les disques durs

Utilisation du **Zoned Bit Recording**



II – Les disques durs

Adressage **C**ylinder/**H**ead/**S**ector vs **L**ogical **B**lock **A**ccess



II – Les disques dur

- Les performances des disques durs sont déterminées par :
  - L'average seek time
  - Le rotatioal speed and latency
  - Media transfert rate

Vitesse	Rotationnal Latency (ms)
5400	5.6
7200	4.2
10000	3.0
15000	2.0

II – Les SSD

- **S**olid **S**tate **D**rive (ou **D**isk)

- Terme utilisé pour de la mémoire flash présentée sous la forme d'un disque (interface, taille, contrôleur, ..)

- Utilisation de mémoire flash NAND
  - 1989 (Toshiba)
  - Problématiques propres à cette technologie « wear leveling

II – Les SSD

Les types de celulles :

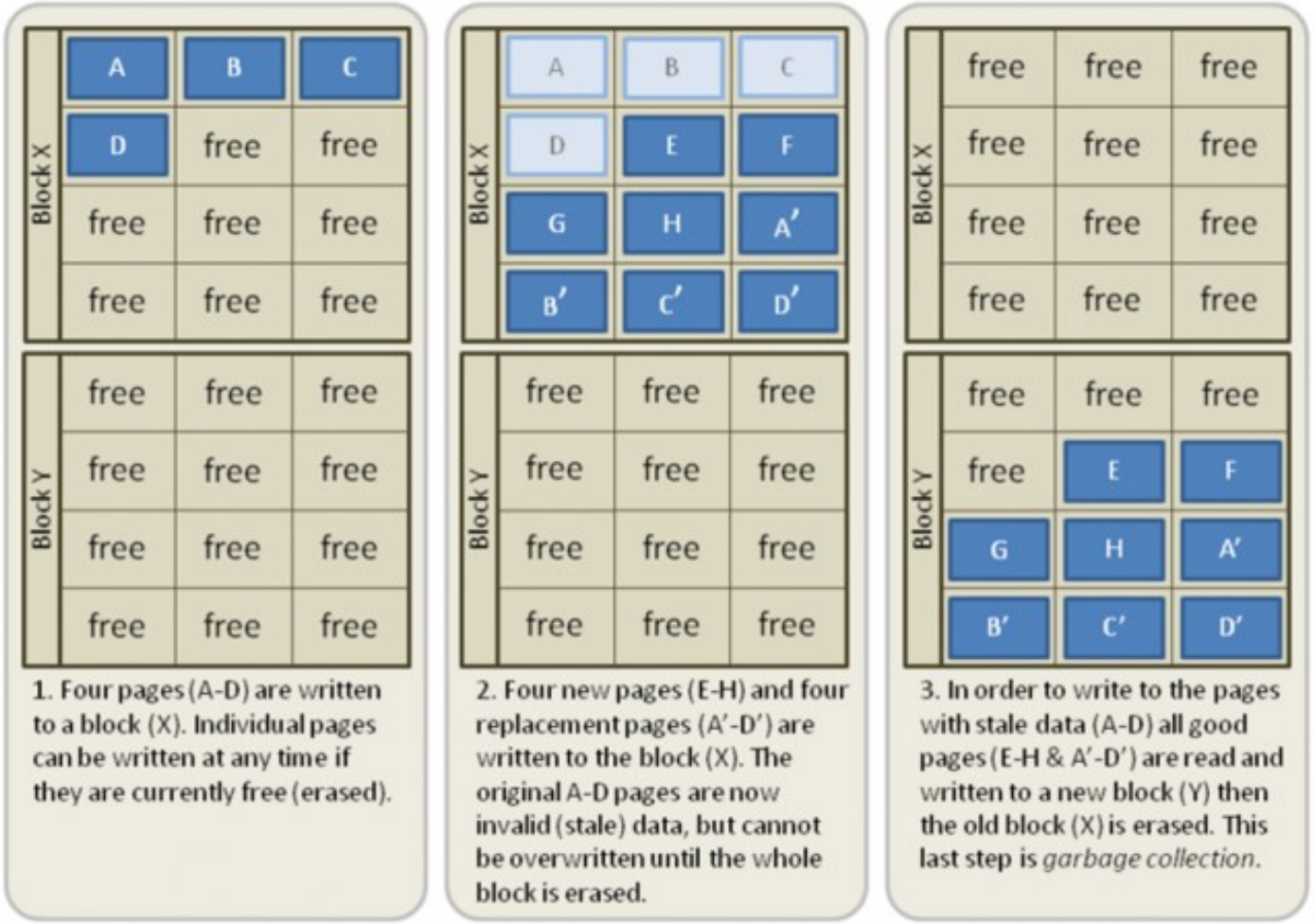
	SLC	MLC	TLC
Bits per cell	1	2	3
P/E Cycles	100,000	3,000	1,000
Read Time	25 µs	50 µs	~75 µs
Program Time	200-300 µs	600-900 µs	~900-1350 µs
Erase Time	1.5-2 ms	3 ms	4.5 ms

Source: http://www.anandtech.com/show/6337/samsung-ssd-840-250gb-review/2

II – Les SSD

- SLC (1 bit per cell) – rapide, couteux
- MLC (2 bits per cell)
  - 3D MLC (2 bits per cell)
- TLC (3 bits per cell)
  - 3D TLC (3 bits per cell)
- QLC (4 bits per cell) – lent, bon marché

II – Les SSD


Block X
A
B
C
D
free
free
free
free
free
free
free
free
free
Block Y
free
free
free
free
free
free
free
free
free
free
free
free
free
free
free
free
1. Four pages (A-D) are written to a block (X). Individual pages can be written at any time if they are currently free (erased).
Block X
A
B
C
D
E
F
G
H
A′
B′
C′
D′
Block Y
free
free
free
free
free
free
free
free
free
free
free
free
2. Four new pages (E-H) and four replacement pages (A′-D′) are written to the block (X). The original A-D pages are now invalid (stale) data, but cannot be overwritten until the whole block is erased.
Block X
free
free
free
free
free
free
free
free
free
free
free
free
Block Y
free
free
free
free
E
F
G
H
A′
B′
C′
D′
3. In order to write to the pages with stale data (A-D) all good pages (E-H & A′-D′) are read and written to a new block (Y) then the old block (X) is erased. This last step is garbage collection.

II – Les SSD

- Les risques propres aux SDD :
  - Le wear leveling
    - Distribution des écritures pour maximiser la vie des cellules, problème : données qui ne sont pas modifiées
    - Cellule de secours
    -
  - Write amplification
    - Utilisation des instructions TRIM
  - Le voltage : plus le voltage est élevé, plus le SSD est « rapide » au détriment de sa durée de vie

II – Les SSD

- Les types d’interfaces physiques utilisées :
  - SAS
  - SATA
  - PCIe

- Les types d’interfaces logiques :
  - ATAPI (SATA) , AHCI (SATA), NVM Express (PCIe)

II – Les SSD

Summary Performance Data – HDD, SSHD, SSD									
Class	Type	FOB IOPS	IOPS (higher is better)			Throughput (larger is better)		Response Time (faster is better)	
Storage Device	Form Factor, Capacity, Cache	RND 4KiB 100% WV	RND 4KiB 100% WV	RND 4KiB 65:35 RWV	RND 4KiB 100% R	SEQ 1024KiB 100% WV	SEQ 1024KiB 100% R	RND 4KiB 100% WV AVE	RND 4KiB 100% WV MAX
HDD & SSHD									
7,200 RPM SATA Hybrid R30-4	2.5" SATA 500 GB WCD	125	147	150	135	97 MB/s	99 MB	15.55 msec	44.84 msec
15,000 RPM SAS HDD IN-1117	2.5" SAS 80 GB WCD	350	340	398	401	84 MB/s	90 MB/s	5.39 msec	97.28 msec
Client SSDs									
mSATA SSD R32-336	mSATA 32 GB WCD	18,000	838	1,318	52,793	79 MB/s	529 MB/s	1.39 msec	75.57 msec
SATA3 SSD IN8-1025	SATA3 256GB WCD	56,986	3,147	3,779	29,876	240 MB/s	400 MB/s	0.51 msec	1,218.45 msec
SATA3 SSD R30-5148	SATA3 256GB WCE	60,090	60,302	41,045	40,686	249 MB/s	386 MB/s	0.35 msec	17.83 msec
Enterprise SSDs									
Enterprise SAS SSD R1-2288	SAS 400GB WCD	61,929	24,848	29,863	53,942	393 MB/s	496 MB/s	0.05 msec	19.60 msec
Server PCIe SSD IN1-1727	PCIe 320GB WCD	133,560	73,008	53,797	54,327	663 MB/s	772 MB/s	0.05 msec	12.60 msec
Server PCIe SSD IN24-1349	PCIe 700GB WCD	417,469	202,929	411,390	684,284	1,343 MB/s	2,053 MB/s	0.03 msec	0.58 msec

II – Les sous systèmes de stockage

- Par sous système de stockage, on entend, le regroupement de :
  - périphériques de stockages
  - contrôleur de stockage
  - de la connectivité réseau
  - de la connectivité pour les périphériques de stockage

- Schéma

II – Les sous systèmes de stockage

- Les sous systèmes sont capables de :
  - Gérer la redondance des données
  - Agréger de la volumétrie
  - Gérer l’ajout/retrait à chaud
  - Gérer l’alimentation électrique
  - Gérer la communication unifiée

II – Les sous systèmes de stockage

- Tiering :
  - Automatique/manuel
  - Mise en cache ou déplacement de données

- Caching
  - Cache Hit/Cache Miss
  - Write-through : écriture sur disque avant validation de l'opération
  - Write-back : validation de l'opération dés réception dans le cache

II – Les sous systèmes de stockage

- L’espace d’adressage présenté aux systèmes distants est appelé « stockage exporté ».

- Au niveau SCSI, un espace d’adressage est une **L**ogical **U**nit, elle est en charge du traitement des commandes SCSI pour des ressources de stockage virtuelles ou physiques

- **L**ogical **U**nit **N**umber
  - C’est un « point d’entrée » pour accéder à des périphériques de stockage (LU, robots, ..)
  - Un LUN ≠ a une ressource exportée
  - Un LUN correspond à une adresse
  - Une LU peut être accédée via plusieurs LUN

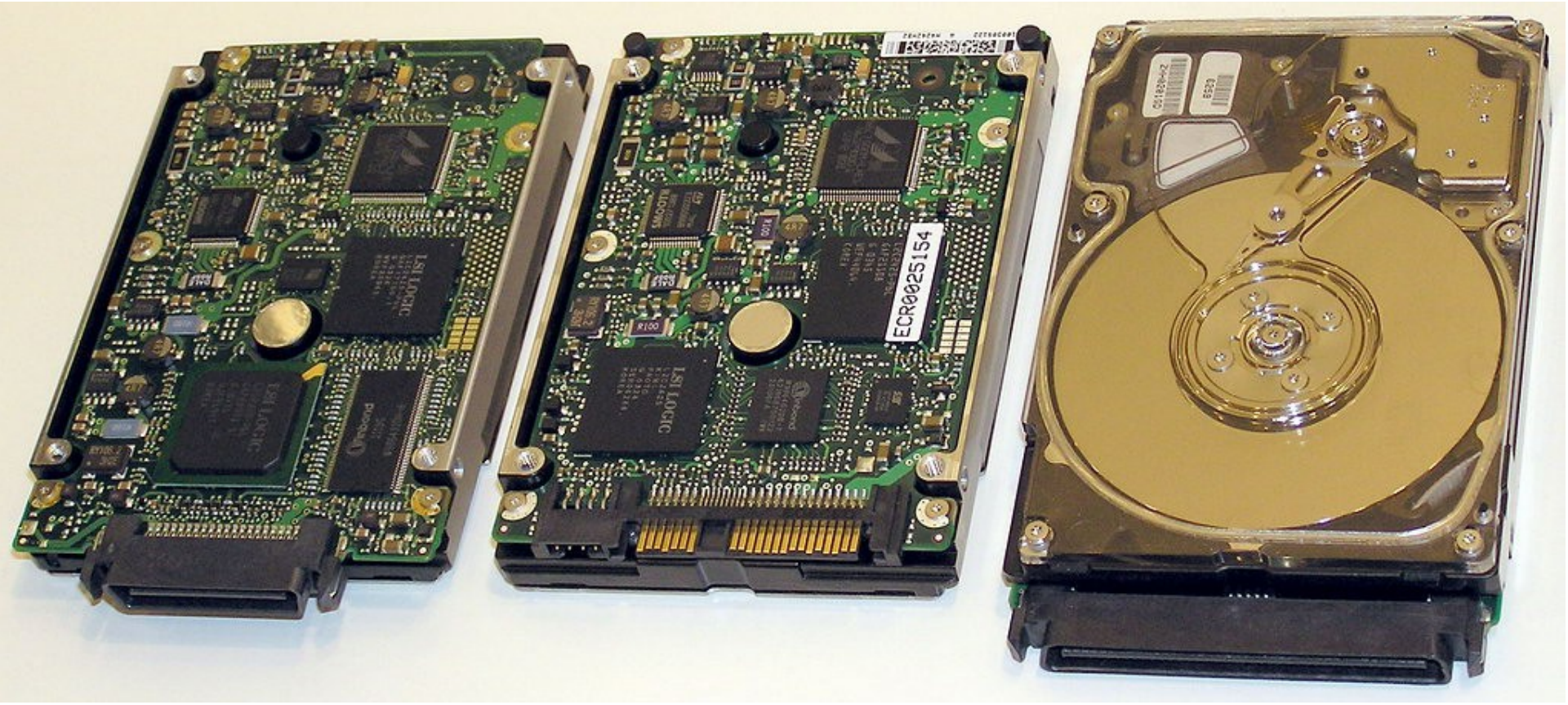
II – SCSI

- **S**mall **C**omputer **S**ystems **I**nterface
- Protocole créé en 1981
- Terme utilisé pour :
  - Le bus SCSI : SCSI Parallel interface
  - Les commandes SCSI
- Au début, forte adhérence entre SCSI (stocker) et parallel SCSI (connecter)
- Le **P**rotocol **D**ata **U**nit utilisé par SCSI est appelé **C**ommand **D**escriptor **B**lock

II – ATA

- **S**mall **C**omputer **S**ystems **I**nterface
- Protocole créé en 1981
- Terme utilisé pour :
  - Le bus ATA : Parallel ATA
  - Les commandes ATA
- Au début, forte adhérence entre SCSI (stocker) et parallel SCSI (connecter)
- Le **P**rotocol **D**ata **U**nit utilisé par SCSI est appelé **C**ommand **D**escriptor **B**lock

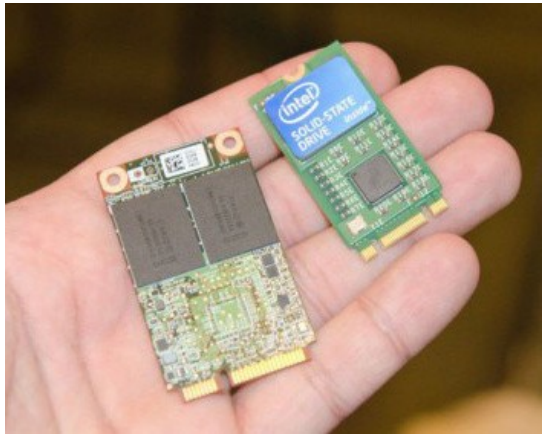
II – Technologies d’interconnexion des périphériques



Fiber Channel AL

SATA

Parallel SCSI



M.
2

PCI Express

IDE

mSATA

II – Technologies d'interconnexion des périphériques (bus)

- **S**CSI **P**arallel **I**nterface
  - Utilisé sur les anciennes machines
  - Utilise les commandes SCSI
  - Début maximum : 640MB/s

- **F**iber**C**hannel-**A**rbitrated**L**oop
  - Commence à disparaitre
  - Très utilisé sur les SAN de type entreprise (latence, vitesse, fonctions)
  - Utilise les commandes SCSI
  - Débit maximum : 16Gbit/s

II – Technologies d'interconnexion des périphériques (bus)

- **S**erial **A**ttached **S**CSI
  - Remplaçant de FC pour connectivité disque/contrôleur
  - Utilise les commandes SCSI
  - Débit maximum : 6-12Gbit/s

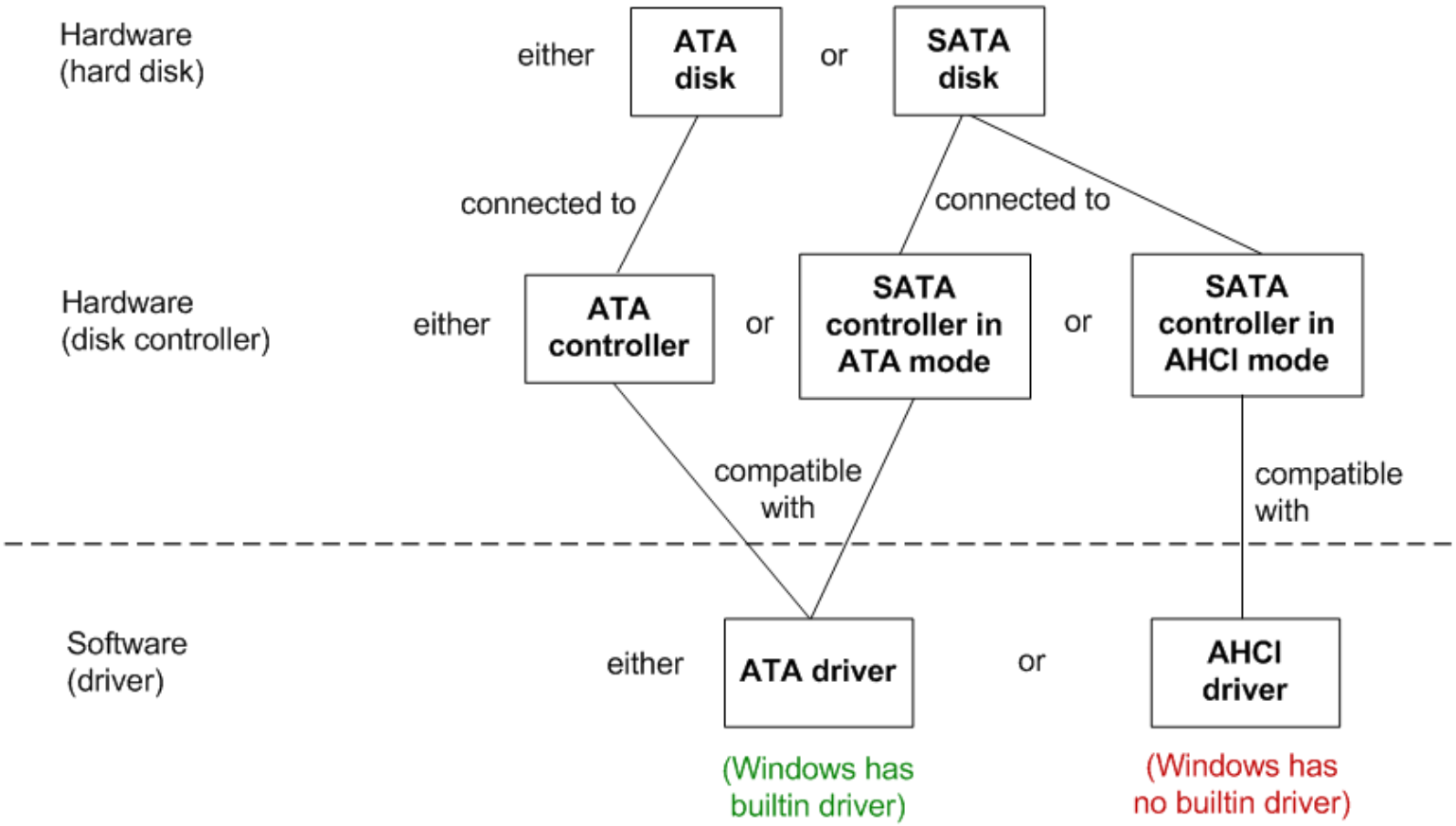
- **N**ear**L**ine-**SAS**
  - Même mécanique que les disques SATA
  - Utilise les commandes SCSI
  - Intelligence des disques SAS (NCQ, double attachement)

II – Technologies d’interconnexion des périphériques (bus)

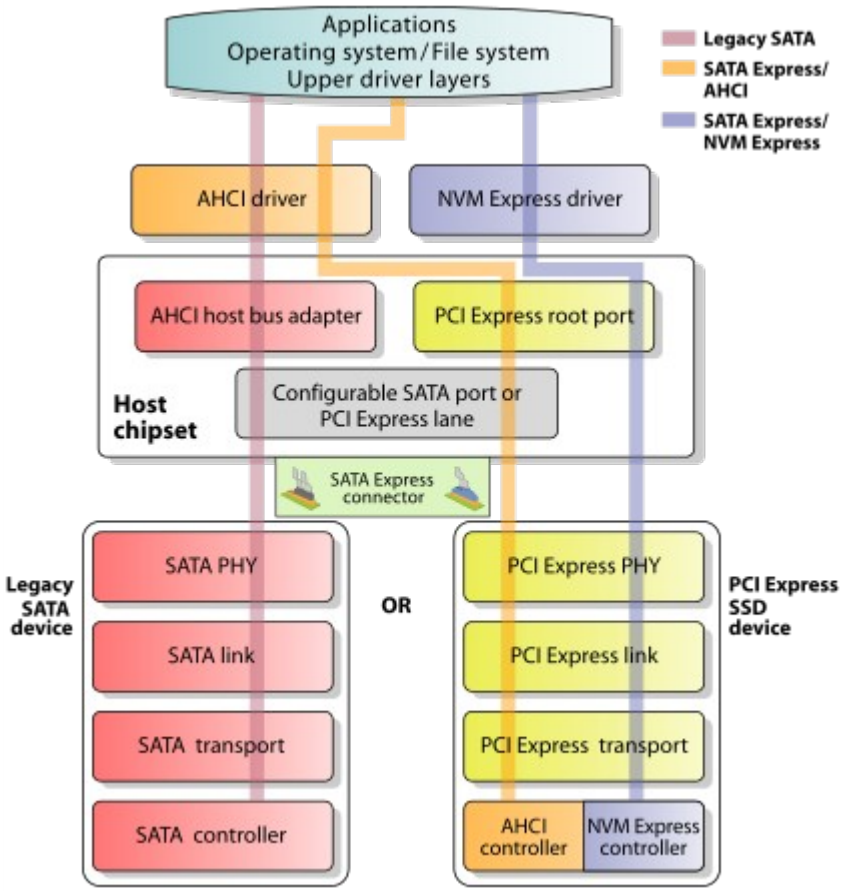
- **AT A**ttachment Interface
  - Aussi nommé IDE
  - Parrallel ATA
  - Standard ATA / ATAPI (commandes ATA)
  - Débit maximum : 133 MB/s (Ultra ATA 133)

- **S**erial **ATA**
  - Evolution d’ATA
  - Supporte les commandes ATA mais possède plus de fonctionnalités (hotplug, NCQ)
  - Débit maximum : 1,5-16Gbit/s
  - Différents connecteurs : mSATA, M.2, mSATA

ATA, AHCI


Hardware
(hard disk)
either
ATA
disk
or
SATA
disk
connected to
connected to
Hardware
(disk controller)
either
ATA
controller
or
SATA
controller in
ATA mode
or
SATA
controller in
AHCI mode
compatible
with
compatible
with
Software
(driver)
either
ATA driver
or
AHCI
driver
(Windows has
builtin driver)
(Windows has
no builtin driver)

AHCI, NVMe



	AHCI	NVMe
Maximum queue depth	One command queue; 32 commands per queue	65535 queues;[29] 65536 commands per queue
Uncacheable register accesses (2000 cycles each)	Six per non-queued command; nine per queued command	Two per command
MSI-X and interrupt steering	A single interrupt; no steering	2048 MSI-X interrupts
Parallelism and multiple threads	Requires synchronization lock to issue a command	No locking
Efficiency for 4 KB commands	Command parameters require two serialized host DRAM fetches	Gets command parameters in one 64-byte fetch

III - Gestion de la redondance des données

- Les menaces
- Les métriques
- Le RAID

III – Les menaces

- Perte de données
  - Suppression ou réécriture de données, destruction de l'équipement de stockage
- Perte d'accès
  - Blocage d'accès sur un composant ou un ensemble de composants
- Perte d'intégrité
  - Modification non voulue des données (bugs, pannes matérielles, erreurs humaines, ..)

III – Les métriques

- MTBF : Mean Time Between Failures
- MTDA : Mean Time between Data Access
- MTDL : Mean Time to Data Lost
- MTTR : Mean Time To Repair

III – Redondance des données

- Duplication
  - On créer une copie supplémentaire de la donnée sur une espace de stockage différent
  - Permet de palier la perte d'accès

- Parité
  - On utilise des systèmes d'encodage pour récupérer des données sans avoir une copie complète des données
  - Permet de palier à la perte d'intégrité

- Delta
  - On « trace » les modifications sur les données
  - Principalement utilisé pour la sauvegarde
  - Permet de palier à la perte de données/et ou d'intégrité

III – Le RAID

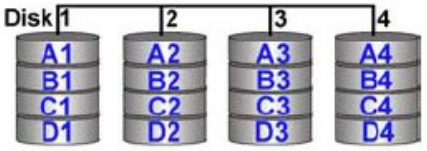
- **R**edundant **A**rray of **I**nexpensive **D**isks
- Créé dans les années 80 à Berkeley pour avoir le même niveau de disponibilité des données que sur les Mainframe
- Un RAID est composé de « membres » regroupés en « ensembles »
- Le RAID se base sur la parité pour assurer l'intégrité des données
- Quand un membre tombe en panne, on parle de fonctionnement en mode dégradé de l'ensemble

III – Le RAID

- Rappels sur le RAID :
  - RAID 0 :
    - Rapide mais fragile
  - RAID 1, 1+0, 0+1
    - Niveau de sécurité correct mais gaspillage d'espace
  - Parity RAID (RAID 5, 6)
    - Bon compromis espace/performance
    - 4-12 disques : RAID 5
    - 12-24 disques : préférer le RAID 6
  - RAID DP (évolution du RAID 4)
    - Plus rapide et sécurisé que du RAID 4

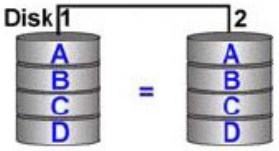
III – Le RAID

RAID 0



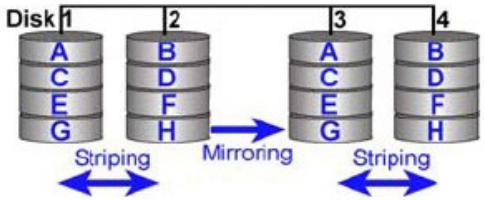
Requires a minimum of one drive.

RAID 1



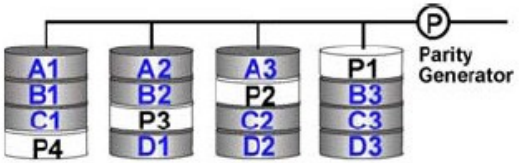
Requires a minimum of two drives.

RAID 1+0



Requires a minimum of four drives.

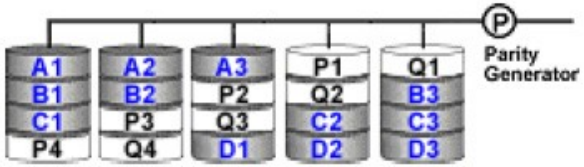
RAID 5



Requires a minimum of three drives.

Pn represents one set of parity.

RAID 6 (ADG)



Requires a minimum of four drives.

Pn and *Qn* represent two sets of parity.

myTOYS

III – Le RAID

- Exemple :
  - Disque Seagate Cheetah 600GB -15K
  - Average latency : 3,65 ms
  - Average seek time : 2 ms
    - 1/((3,65/1000) + (2/1000) = ~175 IOps

- Les pénalités du RAID dans le calcul des IOps:

Type de RAID	Read	Write
RAID 0	1	1
RAID 1 et 10	1	2
RAID 5	1	4
RAID 6	1	6
RAID DP	1	2

III – Le RAID

- Exemple 1:
  - RAID 5 de 10 disques SAS à 15K
  - 50% d'écriture, 50% de lecture
  - Performances brutes :
    - 160* 10 = 1600 IOps
  - Performances « réelles »
    - (1600 x 0,5)/4 +(1600 x 0,5) = 1000 IOps

- Exemple 2 :
  - RAID DP de 12 disques SAS à 15K
  - 20% d'écriture, 80% de lecture
  - Performances brutes :
    - 160* 12 = 1920 IOps
  - Performances « réelles »
    - (1920 x 0,2)/2 +(1920 x 0,8) = 1728 IOps

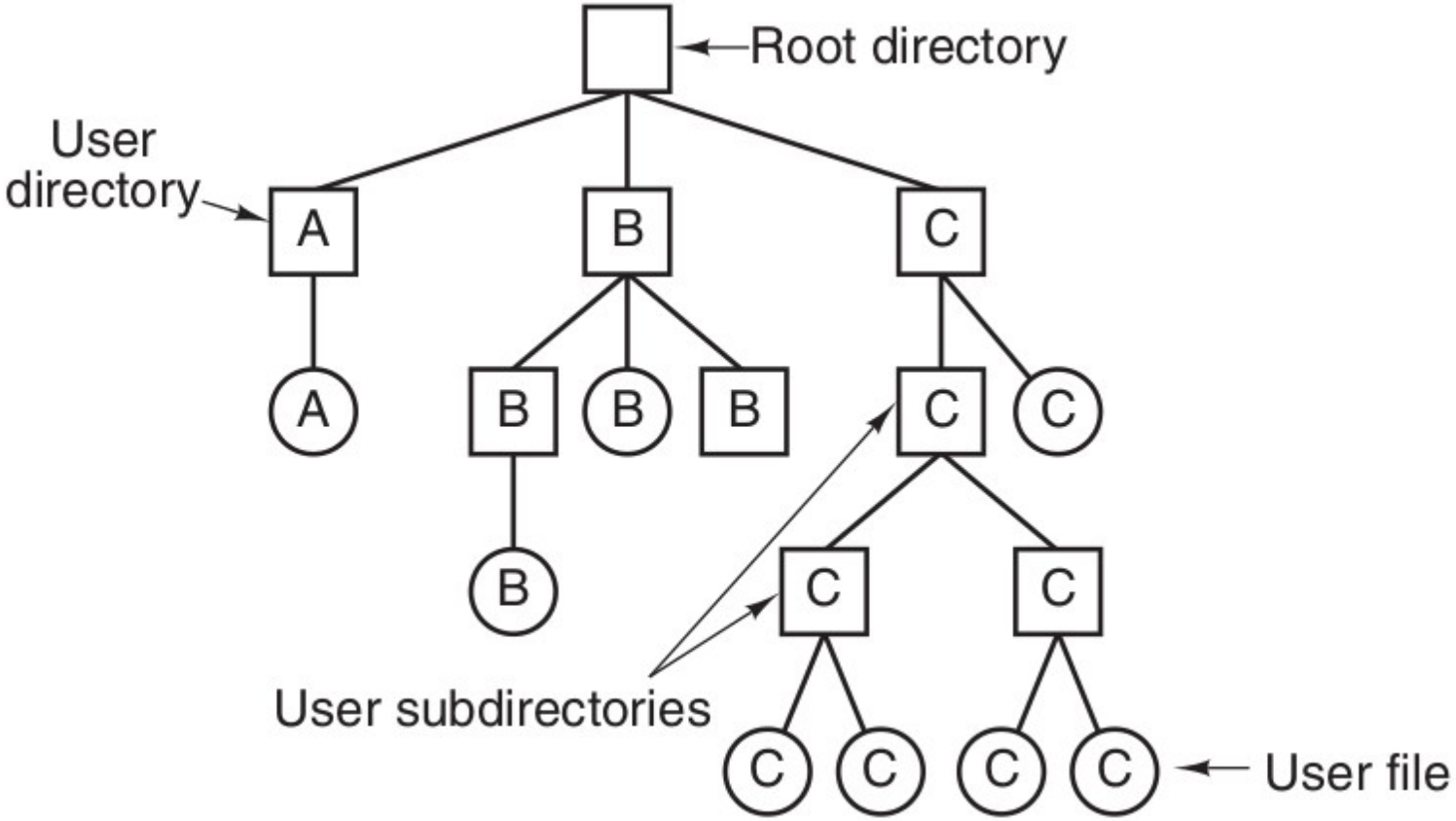
IV - Accès et stockage des données

- Les systèmes de fichiers
- Les NAS

IV – Les systèmes de fichiers

- La taille de la mémoire volatile est limitée
- Des processus différents ont parfois besoin d'accéder à la même donnée
- Il faut stocker les informations quand un processus se termine

IV – Les systèmes de fichiers



IV – Les systèmes de fichiers

- On utilise les fichiers pour créer un niveau d'abstraction et stocker les informations que doivent traiter les processus
- Les fichiers peuvent être considérés comme des espaces d'adressage
- L'objectif d'un FS et de mettre ces fichiers à disposition des utilisateurs et des processus

IV – Les systèmes de fichiers

- Il y a plusieurs types de fichiers :
  - ASCII (on peut les lire en « clair »)
  - Binaire (Il faut en connaître la structure pour les lire)
  - Bloc (Linux)
  - ..

IV – Les systèmes de fichiers

- Un processus peut demander à l'OS l'ensemble du contenu d'un fichier() ou une partie()

IV – Les systèmes de fichiers

- Le FS gère :
  - Des fichiers
  - Des répertoires (qui sont des fichiers spéciaux)
  - Des règles de nommage
  - Des règles d’accès
  - Des attributs (metadata)
  - Des types d’opération (créer, supprimer, lire, ..)
  - Une hiérarchie

IV – Les systèmes de fichiers

- Pour l’installation d’un FS, on parle de formatage.
- Le formatage est réalisé sur un ensemble de blocs (« physiques »)
- On peut utiliser des partitions pour diviser l’espace global en sous espaces contenant chacun un FS différent

IV – Les systèmes de fichiers

- Les FS gère des « logical blocks » qui peuvent agréger plusieurs « physical blocks »

IV – Les systèmes de fichiers

- Les NAS utilisent un système de fichiers qu'ils exportent au client via un protocole (NFS, SMB, ..)
  - Exemple avec NFS :